Attorney's Docket No. K&A 21-0853

APPLICATION

FOR UNITED STATES LETTERS PATENT

SPECIFICATION

TO ALL WHOM IT MAY CONCERN:

BE IT KNOWN THAT I, **WILLIAM J. BUSHEE**, a citizen of UNITED STATES OF AMERICA, has invented a new and useful **AUTOMATIC SYSTEM FOR CONFIGURING TO DYNAMIC DATABASE SEARCH FORMS** of which the following is a specification:

AUTOMATIC SYSTEM FOR CONFIGURING TO DYNAMIC DATABASE SEARCH FORMS

BACKGROUND OF THE INVENTION

Incorporation by Reference

This patent application discloses an invention which may optionally form a portion of a larger system. Other portions of the larger system are disclosed and described in the following co-pending patent applications, all of which are subject to an obligation of assignment to the same person. The disclosures of these applications are herein incorporated by reference in their entireties.

> SYSTEM FOR AUTOMATICALLY CATEGORIZING CONTENT IN HIERARCHICHAL SUBJECT STRUCTURES, Thomas W. Tiahrt, Michael K. Bergman, and William J. Bushee, Filed July ___, 2001, Application Serial Number _____.

Field of the Invention

The present invention relates to network based forms and more particularly pertains to a new automatic system for configuring to dynamic database search forms for facilitate the efficient submission of multiple queries to search engines.

Description of the Prior Art

The Internet is a worldwide system of computer networks in which users at any one computer can get information located on any other computer (given permission). The Internet uses a set of protocols called Transmission Control Protocol/Internet Protocol or TCP/IP. The World Wide Web (often abbreviated as WWW) is a portion of the Internet using hypertext as a method for instant cross-referencing linking one document or site to another.

A database is a collection of data, which is organized in a manner that allows its contents to be easily accessed, managed, and updated. Given this definition an Internet site can be viewed as a database with a collection of data that can be viewed as pages, or accessible documents. Similarly, any network for accessing documents can be considered a database, including intranets and extranets. These network databases can be either static or dynamic. A static network database provides the same set of documents or pages to every user. A dynamic network database presents unique documents or pages in response to a user's query.

The use of search engines using network based forms to find relevant information on the Internet is known in the prior art. Search engines are able to match a multitude of documents to a user's query, but every search engine has a unique layout or form for collecting the users query. Additionally, some sites on the Internet allow additional databases to be searched when a query is applied to the specific form on the site that has been developed for searching the database associated with that particular site. The many variations possible between query form formats on Internet sites has made searching these various search engines and site

databases en masse (or even just in large numbers) difficult if not virtually impossible to perform in an efficient manner. Conventional practice has often involved individually and manually configuring, or setting up, individual query formats suitable for use with queries submitted to each particular search engine or database. This manual configuration technique is in itself is a time consuming, tedious, and not always effective task when performed manually.

However, when one considers the many additional search engines and databases that may be added to the Internet on a daily basis, as well as those existing databases that may be newly discovered, it becomes apparent that this configuring task is an ongoing and never-ending chore that must be repeated for additional search engines and databases if one is to attempt to offer a user a truly in depth search of the ever expanding number of databases. The task of manually configuring search query formats thus hinders and may act as a considerable disincentive to offering a user the capability of searching a full range of the available search engines and databases, and so the user is presented with a less complete search of the Internet than the user may desire.

Therefore, while the prior art systems and techniques may fulfill their respective, particular objectives and requirements, the aforementioned systems and techniques do not disclose a new automatic system for configuring to dynamic database search forms.

SUMMARY OF THE INVENTION

In view of the foregoing disadvantages inherent in the known types of network based forms now present in the prior art, the present invention provides a new automatic system for configuring

to dynamic database search forms construction wherein the same can be utilized for facilitate the efficient submission of multiple queries to a search engines.

The system of the invention includes a computer system having a storage means for facilitating the retention of dynamic database content, a communications means for performing bi-directional communications with a network; a query input means for receiving a plurality of queries from a user and transferring the plurality of queries to a plurality of databases; an action string module for determining an appropriate data entry window for use in passing a query to the database; a results module for locating areas on a responsive page where results are placed; a next link module used to locate a link (URL) associated with additional results provided by the database; and an engine file module for storing results such that a general format query is translatable into a database specific dialects.

The system of the invention generally permits the configuration of a user's query into a number of specific query formats each tailored to the particular requirements of a particular search engine or database. Significantly, the specific query formats are developed in a manner by the system of the invention that requires little, if any, manual or specific intervention by an operator in the development process of each of the specific query formats. The specific query formats may be stored and used for a multitude of user queries.

There has thus been outlined, rather broadly, the more important features of the invention in order that the detailed description thereof that follows may be better understood, and in

order that the present contribution to the art may be better appreciated. There are additional features of the invention that will be described hereinafter and which will form the subject matter of the claims appended hereto.

In this respect, before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and to the arrangements of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments and of being practiced and carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein are for the purpose of description and should not be regarded as limiting.

As such, those skilled in the art will appreciate that the conception, upon which this disclosure is based, may readily be utilized as a basis for the designing of other structures, methods and systems for carrying out the several purposes of the present invention. It is important, therefore, that the claims be regarded as including such equivalent constructions insofar as they do not depart from the spirit and scope of the present invention.

The objects of the invention, along with the various features of novelty that characterize the invention, are pointed out with particularity in the claims annexed to and forming a part of this disclosure. For a better understanding of the invention, its operating advantages and the specific objects attained by its uses, reference should be made to the accompanying drawings and descriptive matter in which there are illustrated preferred embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be better understood and objects other than those set forth above will become apparent when consideration is given to the following detailed description thereof. Such description makes reference to the annexed drawings wherein:

Figure 1 is a schematic functional block diagram of a new automatic system and method for configuring to dynamic database search forms according to the present invention.

Figure 2 is a schematic flow diagram of the next link module of the present invention.

Figure 3 is a schematic flow diagram of the action string module of the present invention.

Figure 4 is a schematic diagrammatic representation of the operation of the results module of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

With reference now to the drawings, and in particular to Figures 1 through 4 thereof, a new automatic system and method for configuring to dynamic database search forms embodying the principles and concepts of the present invention will be described.

As best illustrated in Figures 1 through 4, the automatic system for configuring to dynamic database search forms generally comprises a computer system 20, a query input means 26, and a plurality of function modules 40.

The computer system 20 includes a storage means 22 for

facilitating the retention and recall of dynamic database content. The computer system 20 also includes a communications means 24 for performing bi-directional communications between the computer system 20 and a network.

The query input means 26 is for receiving a plurality of queries from a user and transferring the plurality of queries to a plurality of databases 4.

In one embodiment of the invention, the query input means 26 comprises an input module. The input module may comprise, for example, a keyboard, a mouse, a data input device capable of converting action of a user to a machine readable query, a data file transferred as one or more electrical signals to the computer system 20, a data file transferred as one or more optical signal to the computer system 20, a data file written to memory in the computer system 20, and a data file written to a storage medium accessible by the computer system 20.

The plurality of function modules 40 may include an action string module 42, a results module 44, a next link module 46, and an engine file module 48.

The action string module 42 is interfaced with the computer system 20 for determining a format associated with an entry page for a database 4. The action string module 42 is used for determining an appropriate data entry window for use in passing a query to the database 4.

The results module 44 is also interfaced with the computer system 20 and the action string module 42. The results module 44 locates areas on a responsive page returned by the database 4 in

response to the submitted query where results are placed.

Similarly, the next link module 46 is interfaced to the computer system 20, action string module 42, and results module 44. The next link module 46 locates a URL for a link associated with additional results provided by the database 4 in response to the query.

The engine file module 46 is also interfaced to the computer system 20 and every other module 42, 44, 46 for storing results produced by each module 42, 44, 46 such that a general format query is translatable into a database specific dialect (format) allowing a common query to be submitted to multiple databases 4 each requiring different formats.

A data comparison portion 60 provides user specific information to each of the modules 42,44,46,48 for facilitating analysis of the databases 4. The data comparison portion 60 may include a database listing 62, a bad action string listing 64, a desirable text listing 66, an undesirable text listing 68, an undesirable value listing 70, and a next term listing 72.

The database listing 62 provides a plurality of URLs, each associated with one of the databases 4 to be analyzed.

The bad action string listing 64 provides URLs for known databases (search engines) 4 which are not to be included in the analysis of said databases 4.

The desirable text link listing 66 provides a plurality of desirable terms for use in analysis of the databases 4. The presence of any one of the plurality of desirable terms increases a score

associated with a data entry window (form) on one of the responsive pages.

The undesirable text link listing 68 provides a plurality of undesirable terms for use in analysis of the databases 4. The presence of any one of the plurality of undesirable terms sets a score associated with a data entry window (form) on one of the responsive pages to 0 and ends the analysis of the data entry window.

The undesirable value listing 70 provides a plurality of undesirable values for use in analysis of the databases 4. The presence of any one of the plurality of undesirable values decreases a score associated with a data entry window (form) on one of the responsive pages.

The next link listing 72 provides the next link module 46 with a plurality of candidate terms for facilitating selection of a URL associated with a link to additional responses provided by the database 4 in response to the user's query.

The automatic system for configuring query formats to dynamic database content performs a number of major functions, each of which has multiple sub-functions. The major functions include defining the format of a given database or web page, separating candidate results from all other information returned by the database or web page, finding the location of a link to additional search results (commonly referred to as a "next" button), and writing and verifying an output file specifying the format of the specific database or web page being configured.

Throughout each of its major functions of the system uses a

soft matching technique to improve the probabilities of matching a data-entry window or form to a particular function by approximate term matching. A parameter called a binding factor is used to limit a range of length variations in matching terms. This soft matching technique allows a known term to be matched with similar terms of differing lengths.

An illustrative example of the soft matching technique may be found in the case where a web page is being examined to determine the location of the field on the web page where a request is made for retrieval of the next set of results from a previous search executed on the database of web site. In this illustrative case, the known term "next" may be matched against labels in a web page to find the field, but it is recognized that not all web pages will universally use the term "next" in association with the desired field, and some web pages may use alternate versions of the "next" term. "Next" is made up of four alphabetic characters. If a binding factor of 0 was used during the search, only labels of four characters could be considered. If a binding factor of 1 was used during the search, labels of at least 4 characters, but up to 50% more or 6 characters could be considered. Finally, if a binding factor of two was used during the search, labels of at least four characters but up to 100% more or eight characters could be considered. The following table describes multiple outcomes for different scenarios for this soft matching process.

Label Being Compared to "Next"	B.F. = 0 (4 Characters)	B.F. = 1 (6 Characters)	B.F. = 2 (8 Characters)
NEXT	Match	Match	Match
NEXT O	Fail	Match	Match

NEXT ON	Fail	Fail	Match
NEXT ONE	Fail	Fail	Match
NEXT ONE H	Fail	Fail	Fail
A NEXT	Fail	Match	Match
NE	Fail	Fail	Fail

The first major function of defining the format of a given database or web page initially begins with a URL being provided to the system for an entry page for the database or web page. The system follows the URL and captures the resulting page as a source version. The source version is filtered into a listing of other URLs and text. The text is examined to find forms on the page. These forms can be found in the source version of the entry page by locating each occurrence of the label "form". Each of the forms is scored to determine the most likely candidate for a data entry window for submitting a user's query to be searched, by the database or website. The form with the highest score is then selected as the data entry window for user's queries. If multiple forms have the same score, and that score is the highest score, then the form occurring first on the page is selected.

The database or web site is then tested to determine the most likely position in which results of a user's query will be returned. This information, along with the necessary action strings to drive the query data entry window and the location of items known to not be results, are stored in an engine file.

The invention preferably uses a scoring methodology to develop a numerical representation of a likelihood that any particular data entry window or form is a proper window for submission of a user's query to the database being configured. The

scoring methodology may look at available information to develop multiple numerical metrics used during the evaluation of each data entry window or form.

In general, the metrics are based on web page characteristics (such as text) that are associated with the data entry windows or forms. The numerical metrics may include a bad action string metric, an undesirable value metric, an undesirable link text metric, a desirable link text metric, a name matching metric, a null text metric, and an adjustment metric based upon the number of edit boxes.

In a preferred embodiment, the undesirable link text metric is determined by the presence of an undesirable label associated with the data entry window. These undesirable labels are generally of the type of labels that are associated with data entry windows that are unlikely to lead to suitable databases for the purposes of gathering information. These undesirable labels may preferably include: "email", "horoscope", "weather", "subscribe", "login", "vote", "quote", "update", "remove", "sign up", and "help". If any of these terms are present in the label associated with the data entry window then the scoring operation is halted, and the score for that particular data entry window is set to zero.

Similarly, in a preferred embodiment, the desirable link text metric is determined by the presence of a desirable label associated with the data entry window. These desirable labels are generally of the type of labels that are associated with data entry windows that are highly likely to lead to suitable databases for the purposes of gathering information. These desirable labels may preferably include: "search", "find", "locate", and "query".

Also in a preferred embodiment, the undesirable value metric is determined by the presence of an undesirable value associated with the data entry window. These undesirable labels preferably include: "password" and "zip".

The bad action string metric is a special case that allows elimination of links to known and commonly found search engines which may be found on web pages. If one of these known search engines is listed in the URL associated with the data entry window then the scoring operation is halted, and the score for that particular data entry window is set to zero. Illustrative examples of these known search engines may include, but are not limited to: AltaVista, Google, Amazon, and Yahoo.

The name matching metric is developed by comparing the URL associated with the data entry window with the URL associated with the page to verify that the host name matches on both URLs.

If no button text is associated with a data entry window or form, the scoring methodology sets the null text metric to a non-zero value to prevent the form from being unnecessarily discriminated against if the form is implemented as an image button (for example, an icon) with no text associated with the form.

In a preferred embodiment, multiple binding factors are used to facilitate soft matching of the text, values, URLs, and labels during the scoring operation. The scoring operation may use a binding factor of 0 associated with the bad action string metric, a binding factor of 1 associated with the undesirable link text metric, a binding factor of 2 associated with the desirable link text metric, and a binding factor of 0 associated with the undesirable values

metric.

Also in a preferred embodiment, each one of the metrics may be assigned a unique integer value based on the validity of that particular metric. The magnitude of the integer value may be determined by the relative importance of the metric as compared to all other metrics. Illustratively, the scoring methodology may assign an integer value of 6 for any data entry window that has a URL host name that matches the host name of the web page; or an integer value of 0 for any data entry window which does not have a URL host name which matches the host name of the web page. The scoring methodology may also assign an integer value of 4 for any data entry window that has desirable link text; or an integer value of 0 for any data entry window that does not have desirable link text. Similarly, the scoring methodology may assign an integer value of 4 for any data entry window that does not have undesirable values; or an integer value of 0 for any data window that does have undesirable values. Finally, the scoring methodology may also assign an integer value of 2 for any data entry window that has no associated text.

A numeric score may be computed for each data entry window by adding the integer values associated with each of the metrics, such as the name match metric, the undesirable value metric, the desirable text metric, and the null text metric. The numeric score may then be adjusted by adding an integer value representing the number of data entry windows found versus the number of data entry windows expected. In a preferred embodiment, this adjustment may range from 3 to 0, inclusive.

The next major function of the system is finding the location

of query results. This process is performed by selecting multiple validation queries to be submitted to the database using the data entry window selected by the previous major function (defining the format), submitting the validation queries and analyzing the responsive pages returned by the database in response to the validation queries.

In a preferred embodiment, the terms "home", "copyright", and "web" are used as validation queries. In a further preferred embodiment, three terms common to the subject area, with minimal overlap between any two terms, may be substituted as validation queries.

The first validation query is submitted twice and the responsive pages associated with each submittal are captured and analyzed. Any difference between these two pages indicates areas that are not used to return results associated with a query.

Additionally, one of the responsive pages associated with the first validation query is compared to both of the responsive pages associated with the second and third validation queries. Any commonalities between these responsive pages also indicate areas that are not used to return results associated with a query.

Finally, the responsive page associated with the second validation query and the responsive page associated with the third validation query are compared. Similarly, any commonalities between these responsive pages also indicate areas that are not used to return results associated with a query.

In each of these steps the comparison is done by examining URLs and text (labels) associated with the URLs. In a preferred

embodiment, a value of 0.1 may be used as a binding factor associated with the text.

The third major function, finding the location of the "Next" link, builds upon the results of the last major function by using the results obtained in response to the validation queries. The text (labels) in the first responsive page is compared sequentially with a listing of candidate terms. A label matching the earliest term from the listing is selected as the "Next" link. If no matches are found, the responsive page captured in association with the second validation query is similarly analyzed. A binding factor may be used during the comparison of the text with the candidate terms. Preferably, the binding factor has a value ranging from 1 to 2 inclusive. Most preferably the binding factor has a value of 1.5.

In a preferred embodiment, the listing of candidate terms may include: "next", "next page", "[next page]", "next results", "next ", "next>>", "next hits", " next>>", ">>", ">>", ">", "&gt", "more results", "next 5", "next 10", "next 15", "next 20", "next 50", "next 100", "weiter", "display next", "get more search results", "more results", "page down", "arrow", "click here for more", "get the next", "go to", "more", "site matches". "more in this category", "search for more", "search for next", "show all", "show more results matching", "siguiente", "nächste", "siguiente pág", "próximos 10 regidtros", "10", "15", "20", "25", and "30".

The final major function of the system is storing the results in a file (such as a form method indicator) which may be used during a search process for automatically translating a query in a general format into each dialect (alternate format) required by each

database or search engine.

The foregoing is considered as illustrative only of the principles of the invention. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the invention to the exact construction and operation shown and described, and accordingly, all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.